Extracción Automática de Nombres Científicos en Documentos Digitales

Alejandro Molina

CONABIO

Dic. 2015

Minería de datos y biodiversidad

- *¿Qué es la minería de datos?* métodos de estadística, matemáticas y computación para encontrar información a partir de datos "no estructurados"
- *¿Para qué se puede aplicar?* para encontrar nombres de especies en colecciones grandes de documentos, entre otras cosas...

Minería de datos en CONABIO

- Biblioteca para minería de textos en biodiversidad (programadores): `https://bitbucket.org/conabio_cmd/text-mining`
- Sistema Web para nombres de especies (biólogos): `http://52.88.164.60:8080/`

Extracción automática de nombres científicos en documentos digitales

En este taller:

- Obtener texto plano a partir de otros formatos
  - PDF
  - WORD
  - JPG
  - ...
- Buscar nombres de especies
  - Recursos taxonómicos
  - Lista propia
  - Contexto (*machine learning*)

¡Manos a la obra!

`http://52.88.164.60:8080/`